Supplementary Methods for:

Integrating Genotypic and Gene Expression Data to Identify Key Drivers of Complex Traits

Eric E. Schadt[1‡], John Lamb[1], Xia Yang[4], Jun Zhu[1], Steve Edwards[1], Debraj GuhaThakurta[1], Solveig K. Sieberts[1], Stephanie Monks[2], Marc Reitman[5], Chunsheng Zhang[1], Pek Yee Lum[1], Amy Leonardson[1], Rolf Thieringer[6], Joseph M. Metzger[7], Liming Yang[7], John Castle[1], Haoyuan Zhu[1], Shera F. Kash[8], Thomas A. Drake[3], Alan Sachs[1], Aldons J. Lusis[4]

[1]Rosetta Inpharmatics, LLC, a wholly owned subsidiary of Merck & Co., Inc., Seattle, WA 98109; [2]Dept. of Statistics, Oklahoma State University, Stillwater OK 74078; [3]Dept. of Pathology and Laboratory Medicine, [4]Dept. of Microbiology, Molecular Genetics, and Immunology, Dept of Medicine, and Dept of Human Genetics, UCLA, Los Angeles CA 90095; [5]Dept. of Metabolic Disoders, [6]Dept. of Cardiovascular Disease, [7]Dept. of Pharmacology, Merck Research Laboratories, Rahway, NJ 07065, and [8]Deltagen, Inc., San Carlos, CA 94070.

Simulating Causal, Reactive and Independent Relationships Among Traits to Assess the Power of the LCMS Procedure. The LCMS procedure, which assesses whether the data support a causal, reactive or independent relationship among traits controlled by the same locus, can be validated in several ways. As a first validation step, we simulated traits under the control of a common locus assuming either an independent or causal/reactive model to assess the power to detect the true model. Supplementary Figure 1 highlights power curves for 5 different simulated models. In each case genotypes for the hypothetical locus were simulated assuming an F2 intercross population of size 360. The genetic map and cross were simulated using the Rmap and Rcross programs in the QTL Cartographer software package[1]. A primary trait, R_P, was then simulated based on a simple additive linear relationship between simulated quantitative trait values and genotypes at the simulated locus, where the strength of association was fixed as indicated by the directed edges linking the locus to R_P in Supplementary Figure 1. The residuals of this simulated relationship between locus and R_P were taken to be normally distributed.

After simulating the locus genotypes and primary trait for each causal model indicated in Supplementary Figure 1, secondary traits were simulated for each model with varying degrees of association with the primary trait. Note that the higher the association with the primary trait, the higher the association with the trait locus genotypes. We also simulated the independence model in which the two traits were conditionally independent given the genotype. For the causal/reactive models the secondary trait cannot achieve a stronger association to the locus than the primary trait, since it depends on the genotype only through the primary trait. Thus, as the genetic correlation of the secondary trait approaches that of the primary trait, the correlation between the traits approaches 100%. The associations were simulated such that the correlations between locus and secondary trait were varied in 0.001 increments from a coefficient of determination (r^2) of 0 up to the maximum possible value determined by the given model. For each such value the secondary trait was simulated 1000 times, and the likelihoods for the 3 possible models for each simulation were fit to the data. The model with an AIC significantly smaller than the AIC's of the competing models was noted for each simulation. The threshold to determine whether an AIC was significantly smaller than competing AIC's was determined empirically by constraining the false positive rate to be less then 5%. In addition, the association between the secondary trait and locus genotypes had to be significant at the 0.05 level before the AIC comparison could be considered valid. The power for each r^2 tested was then taken as the number of times the true model was chosen, divided by 1000, the number of simulations considered for each coefficient of determination value considered. While we use a 0.05 significance cutoff to highlight the power of the LCMS procedure, in practice we choose the model with the lowest AIC. Using this strategy the proportion of time the correct model is chosen when applied to the simulated data is almost always 100%.

For all of the models the power is seen to drop off dramatically as the maximum association between locus genotypes and secondary trait that can be realized by a given model is achieved. For the independent model this drop in power is caused by the secondary trait becoming 100% correlated with the locus genotypes, so that the secondary trait and locus genotypes become indistinguishable. This symmetry results in an inability to discriminate the independent model from the causal/reactive model. The situation is similar for the causal/reactive models, where the primary and secondary traits become 100% correlated as the maximum association possible between the locus genotypes and secondary trait is achieved. Again, this symmetry makes it impossible to discriminate the two traits, so that the model in which the primary trait is causal for the secondary trait is indistinguishable from the model in which the secondary trait is causal for the primary trait.

Linkage Disequilibrium versus Pleiotropic Effects. The test for pleiotropy vs. close linkage described in the main text was tested in the following way. Let Y_1 and Y_2 represent quantitative trait random variables, with QTL Q_1 and Q_2 at positions p_1 and p_2, respectively. The primary hypothesis to test is whether $p_1 = p_2$, indicating a pleiotropic effect at the QTL for traits Y_1 and Y_2. Zeng *et al.*[2] devised statistical tests to assess whether the positions are equal. We have developed a slight generalization of this test. Because the positions under consideration for this test will be relatively close together on a given chromosome (e.g., within 20 cM), we would expect Y_1 and Y_2 to be correlated if the QTL effects at each location are significant enough, and so we form the most basic model for these traits under the control of a single, common QTL as

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} Q + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix},$$

where Q is a categorical random variable indicating the genotypes at the position of interest, and $\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$ is distributed as a bivariate normal random variable with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 \\ \sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix}$.

The case where $p_1 = p_2$ represents the null hypothesis of pleiotropy. The aim is to test this null against a more general alternative hypothesis that indicates $p_1 \neq p_2$. The alternative hypotheses of interest can be captured by the following model:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \beta_1 & \beta_2 \\ \beta_3 & \beta_4 \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix},$$

where the ε_i are distributed as for the pleiotropy model. We are now in a position to test the null hypothesis against any of a series of alternative hypotheses. The likelihoods for the 2 competing models are easily formed, and maximum likelihood methods are then employed to estimate the model parameters (μ_i, β_j, and σ_k). With the maximum likelihood estimates in hand, we can form the likelihood ratio test statistic to directly test the null hypothesis against the alternative.

There are several alternative hypotheses that could be tested in the setting. The most relevant for our purposes is:

1. $H_A : \beta_1 \neq 0, \beta_4 \neq 0, \beta_2 = 0, \beta_3 = 0,$

indicating closely linked QTL with no pleiotropic effects. Other alternative hypotheses that could be tested are:

2. $H_A : \beta_1 \neq 0, \beta_4 \neq 0, \beta_2 \neq 0, \beta_3 = 0,$

indicating closely linked QTL with pleiotropic effects at the first position,

3. $H_A : \beta_1 \neq 0, \beta_4 \neq 0, \beta_2 = 0, \beta_3 \neq 0,$

indicating closely linked QTL with pleiotropic effects at the second position, and

4. $H_A : \beta_1 \neq 0, \beta_4 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0,$

indicating closely linked QTL with pleiotropic effects at both positions. Other null hypotheses and corresponding alternative hypotheses naturally follow from the general models presented here.

Forming the Likelihoods for the LCMS Procedure. Each of the 3 models depicted in Figure 1a induces a particular correlation structure between C and R that can be easily modeled by placing mathematical constraints on the underlying joint probability distribution for the variables (L,C,R). If we assume standard Markov properties for the simple graphs depicted in Figure 1a, then the joint probability distributions for the 3 models of interest are:

M1. $P(L,R,C)=P(L)P(R|L)P(C|R)$

M2. $P(L,R,C)=P(L)P(C|L)P(R|C)$

M3. $P(L,R,C)=P(L)P(R|L)P(C|R,L)$

If we further assume the traits R and C are normally distributed about each genotypic mean at the common locus L, the likelihoods corresponding to each of the joint probability distributions given above are easily constructed, as detailed below. The likelihood-based causality model selection (LCMS) procedure introduced in the main text then consists of selecting the model with the highest likelihood given the data, which is the model that is best supported by the data. Let R be a gene expression trait for some gene g, and let C be a classic quantitative phenotype (e.g., a clinical trait) . Without loss of generality, we treat R and C as quantitative traits (similar arguments would hold for qualitative traits). For the association between R and C, it is of interest to determine those genetic and environmental components driving the association, and it is of interest to determine whether we can assess in a genetics context whether one trait drives the other. That is, does one of the following relationships hold:

$$\mathrm{R} \longrightarrow \mathrm{C} \qquad \mathrm{R} \longleftarrow \mathrm{C}$$

It is not possible to look at these two traits in isolation and determine which, if any, of these cases holds. In the more classical graphical modeling context, where the aim is to reconstruct complex networks of interaction, different graphical structures are assessed and edges are weighted and directed in such structures using conditional mutual information measures that examine, for instance, all adjacent triplets (say, $X, Y,$ and Z) in the graph, where the topology of the graph is constrained a priori to satisfy certain Markov properties. Without the genetic information discussed in the main text, this network reconstruction problem is difficult because many of the different possibilities that are considered are not easily distinguishable[3]. For instance, consider the following three possible relationships among three traits of interest:

i) $X \longrightarrow Y \longrightarrow Z$

ii) $X \longleftarrow Y \longrightarrow Z$

iii) $X \longrightarrow Y \longleftarrow Z$

Here we see that cases i) and ii) are not distinguishable because they have the same dependency structure. This presents problems for reliable reconstruction of genetic networks given correlation data alone, since in many instances it will not be possible to direct edges (directing the edges in such graphs establishes the cause and effect relationships of interest to us in reconstructing pathways associated with disease), and reliably directing the edges can require a considerable amount of data.

In our present case we have a significant advantage given the relationships between gene expression and clinical traits and quantitative trait loci (QTL). The QTL information provides an extremely powerful filter as we are able to restrict attention from all significantly correlated genes and trait values, to those subsets of genes and traits that are under the control of a common set of QTL. Our triplets then become QTL and traits, where we are able to initially direct an edge between the QTL and a single trait by definition of a QTL, and then test all other traits pair wise as discussed below to determine how the trait pairs are positioned relative to one another. For instance, going back to the case where we have a clinical trait C linked to a QTL L, we are able to immediately fix:

$$\mathrm{L} \longrightarrow \mathrm{C}$$

This relationship holds because L is a QTL for C, and the QTL gives us the direction since it is causal for C (i.e., variations in C do not cause variations in the DNA such that the changes give rise to a QTL, rather DNA variations underlying a given QTL lead to variations in C). One method to position a given gene expression trait, R, relative to C, is to test for mutual independence of L and C given R. That is, if C and L are truly independent given R, then we know the (L,C,R) triplet has the form given in M1 in Figure 1a of the main text.

In applying this test to uncover the true relationship between the traits, one can consider all possible correlation structures induced by the different models and use likelihood methods to determine what model is best supported by the data. Towards that end, likelihood models for each of the three cases considered above are constructed. Beginning with the first case, we want to establish whether C is correlated with the genotypes at L, conditional on R, i.e., we want to assess if the following relation among the three variables holds:

$$P(C,L|R)=P(C|R)P(L|R).$$

This conditional probability is related to the mutual information measure that is typically used in network reconstruction problems:

$$I(C,L|R)=\sum_{C,L,R}P(C,L,R)\log\left(\frac{P(C,L|R)}{P(C|R)P(L|C)}\right),$$

where the summation symbol indicates the continuous variables C and R have been discretized to allow for efficient computation over complicated graph structures, as is usually done in network reconstruction problems[3]. While the mutual information measure is useful in more general network reconstruction problems, the problem described here is significantly easier than the general case, and so, leads to a more robust and more powerful test for the purpose of establishing the relationship between any two traits, although, the sort of "test" described here can be systematically applied to reconstruct complex gene networks.

Without loss of generality and given the application here to the BXD data set described in the main text, we assume an F2 population derived from two inbred strains of mice. Beginning with individual animals in an F2 population, the likelihoods associated with each of the component pieces of the joint probability distributions for the 3 models given in the main text follow from the simple regression models:

$$r_i = \mu + \alpha_R f\left(L\right) + \delta_R g\left(L\right) + \varepsilon_R$$

and

$$c_i = \mu + \alpha_C f\left(L\right) + \delta_C g\left(L\right) + \varepsilon_C ,$$

where the r_i and c_i represent the measurements for the expression and clinical traits for individual i in the population, α and δ are the additive and dominance effects, ε_R and ε_C are normally distributed with mean 0 and variance σ_R^2 and σ_C^2, respectively, and the functions f and g are constructed from the genotype probability distribution for locus L as previously described[4]. From this parameterization of the regression models, the likelihoods for an individual animal associated with the joint probabilities given above are

1. $l\left(\theta_{r_i|L};r_i\,|\,L\right)=\frac{1}{\sqrt{2\pi}\sigma_R}\exp\left(-\frac{\left(r_i-\mu_{R_L}\right)^2}{2\sigma_R^2}\right)$, with $\theta_{r_i|L}=\left(\mu_{R_L},\sigma_R\right)$,

2. $l\left(\theta_{c_i|L};c_i\,|\,L\right)=\frac{1}{\sqrt{2\pi}\sigma_C}\exp\left(-\frac{\left(c_i-\mu_{C_L}\right)^2}{2\sigma_C^2}\right)$, with $\theta_{c_i|L}=\left(\mu_{C_L},\sigma_C\right)$,

3. $l\left(\theta_{c_i|r_i};c_i\,|\,r_i\right)=\frac{1}{\sqrt{2\pi\sigma_C^2\left(1-\rho^2\right)}}\exp\left[-\frac{\left(c_i-\mu_C-\rho\frac{\sigma_C}{\sigma_R}\left(r_i-\mu_R\right)\right)^2}{2\sigma_C^2\left(1-\rho^2\right)}\right]$,

   with $\theta_{c_i|r_i}=\left(\mu_R,\mu_C,\sigma_R,\sigma_C,\rho\right)$,

4. $l\left(\theta_{r_i|c_i};r_i\,|\,c_i\right)=\frac{1}{\sqrt{2\pi\sigma_R^2\left(1-\rho^2\right)}}\exp\left[-\frac{\left(r_i-\mu_R-\rho\frac{\sigma_R}{\sigma_C}\left(c_i-\mu_C\right)\right)^2}{2\sigma_R^2\left(1-\rho^2\right)}\right]$,

   with $\theta_{r_i|c_i}=\left(\mu_R,\mu_C,\sigma_R,\sigma_C,\rho\right)$,

5. $l\left(\theta_{c_i|r_i,L};c_i\,|\,r_i,L\right)=\frac{1}{\sqrt{2\pi\sigma_C^2\left(1-\rho^2\right)}}\exp\left[-\frac{\left(c_i-\mu_{C_L}-\rho\frac{\sigma_C}{\sigma_R}\left(r_i-\mu_R\right)\right)^2}{2\sigma_C^2\left(1-\rho^2\right)}\right]$,

   with $\theta_{c_i|r_i,L}=\left(\mu_R,\mu_{C_L},\sigma_R,\sigma_C,\rho\right)$,

6. $l\left(\theta_{r_i|c_i,L};r_i\,|\,c_i,L\right)=\frac{1}{\sqrt{2\pi\sigma_R^2\left(1-\rho^2\right)}}\exp\left[-\frac{\left(r_i-\mu_{R_L}-\rho\frac{\sigma_R}{\sigma_C}\left(c_i-\mu_C\right)\right)^2}{2\sigma_R^2\left(1-\rho^2\right)}\right]$,

   with $\theta_{r_i|c_i,L}=\left(\mu_{R_L},\mu_C,\sigma_R,\sigma_C,\rho\right)$,

where $l\left(\theta_{r_i|L};r_i\,|\,L\right)$, $l\left(\theta_{c_i|L};c_i\,|\,L\right)$, $l\left(\theta_{r_i|c_i};r_i\,|\,c_i\right)$, $l\left(\theta_{c_i|r_i};c_i\,|\,r_i\right)$, $l\left(\theta_{r_i|c_i,L};r_i\,|\,c_i,L\right)$, and $l\left(\theta_{c_i|r_i,L};c_i\,|\,r_i,L\right)$ correspond to $P\left(R|L\right)$, $P\left(C|L\right)$, $P\left(R|C\right)$, $P\left(C|R\right)$, $P\left(R|C,L\right)$, and $P\left(C\,|\,R,L\right)$, respectively. The specific form of μ_{R_L} and μ_{C_L} depends on the locus genotype. For example, given the Falconer parameterization for the trait/QTL regression models given above, μ_{R_L} can take on values $\mu-\alpha_R$, $\mu+\delta_R$, and $\mu+\alpha_R$, depending on which of the 3 possible genotypic states in the F2 population at locus L is under consideration. For each trait we have assumed the distribution about each genotype at locus L has constant variance. Given these forms of the components of the likelihoods for a single animal, the likelihoods for each model over all animals in the population of interest are given by:

1. $L\left(\theta_{M_1};M_1\right)=\prod_{i=1}^{N}\sum_{j=1}^{3}P\left(L_j\right)l\left(\theta_{r_i|L_j};r_i\mid L_j\right)l\left(\theta_{c_i|r_i};c_i\mid r_i\right)$

2. $L\left(\theta_{M_2};M_2\right)=\prod_{i=1}^{N}\sum_{j=1}^{3}P\left(L_j\right)l\left(\theta_{c_i|L_j};c_i\mid L_j\right)l\left(\theta_{r_i|c_i};r_i\mid c_i\right)$

3. $L\left(\theta_{M_3};M_3\right)=\prod_{i=1}^{N}\sum_{j=1}^{3}P\left(L_j\right)l\left(\theta_{r_i|L_j};r_i\mid L_j\right)l\left(\theta_{c_i|r_i,L_j};c_i\mid r_i,L_j\right)$,

where the parameter vectors for each likelihood $\left(\theta_{M_1},\theta_{M_2},\theta_{M_3}\right)$ are taken as the union of the component parameter vectors given above for each of the component pieces making up the model. Here the sum over the L_j represent the 3 possible QTL genotypes that obtain at a given locus in an F2 population constructed from inbred strains of mice. For each likelihood model, the corresponding likelihood is maximized and parameters are estimated using standard maximum likelihood methods. The AICs are then computed for each model as two times the negative of loglikelihood, maximized over the parameters, plus two times the number of parameters. The model associated with the smallest AIC value is identified as that which is best supported by the data.

An alternative to the likelihood approach described above to identify the model best supported by the data involves the use of conditional correlations. For example, if model 1 holds, then the correlation between C and L conditional on R will not be significantly different from *0*. On the other hand, if model 2 holds, then the correlation between R and L conditional on C will not be significantly different from *0*. If the conditional correlations in both cases estimate to be significantly greater than *0*, then we can conclude that model 3 is best supported by the data. Of course, if the conditional correlations both estimate to be statistically indistinguishable from *0*, then we can the results are inconclusive.

Statistical Test for Fat Mass Differences Between Wildtype and Knockout Mice.

Given the experimental design for assessing fat mass differences between wildtype (WT) and knockout (KO) animals involves multiple repeated measures over a number of time points for each animal, we leverage this longitudinal data to enhance the power to detect differences at any given time point using autoregressive models[5]. In particular, if we let y_{tl} represent the fat mass measure for animal l at time point t, then our autoregressive model of interest is

$$y_{tl} = \gamma_{0t} + \gamma_{1t} Q_l + \sum_{j=1}^{t} \phi_{tj} \left(y_{t-j,l} - \hat{y}_{t-j,l} \right) + w_{tl}$$

for $t = 1,2,3,4,5,6,7$ (the 7 time points over which fat mass measures were taken), where $\mathrm{var}(w_{tl}) = \sigma_t^2$, for $l = 1,\ldots,n$, and n is the number of animals in the study. Here, Q_l is the genotype indicator for the gene of interest, taking the value 0 for wildtype animals and 1 for KO animals. $y_{t-j,l} - \hat{y}_{t-j,l}$ represents the fat mass measures at time point $t - j + 1$, where $\hat{y}_{t-j,l}$ is the prediction of $y_{t-j,l}$ from the previous model. In taking this difference as the fat mass measure for a time point of interest, the effect of genotype on fat mass at previous time points is effectively removed.

After fitting the model at the desired time point, the parameter γ_{1t} can be tested for significance using a standard t test. Therefore, the null hypothesis for the test is that there is no difference in fat mass between the KO and WT groups, conditional on genotypic effects from the previous time points.

Construction and phenotyping of Zfp90 transgenic, C3ar1-/-, Tgfbr2+/-, and control mice.

C3AR1-/- mice were obtained from Deltagen, Inc. (San Carlos, CA). A 6.93 kb IRES-lacZ reporter and neomycin resistance cassette (IRES-lacZ-neo) was subcloned into a 5.0 kb fragment isolated from a mouse genomic phage library, such that 197 base pairs coding for the protein were replaced by IRES-LacZ-neo (Supplementary Fig. 3a). The IRES-lacZ-neo cassette was flanked by 3.4 kb of mouse genomic DNA at its 5´ aspect and 1.6 kb of mouse genomic DNA at its 3´ aspect. The targeting vector was linearized and electroporated into 129/OlaHsd mouse embryonic (ES) stem cells. ES cells were selected for G418 resistance, and colonies carrying the homologously integrated neo DNA were identified by PCR amplification using a 5´ neo-specific primer (5´-GGGATCTTGGCCATGGTAAGCTGAT-3´) paired with a primer located outside the targeting homology arms on the 5´ side (GS1: 5´-CAGCATCAAAAGCTGCACAGCGAGG-3´). The homologous recombination event was confirmed on the 3´ side using a 3´ neo-specific primer (5´-ACGTACTCGGATGGAAGCCGGTCTT-3´) paired with a primer located outside the targeting homology arm on the 3´ side. (GS2: 5´-GTGGCATTTGGCACTGTGTTCTGTC-3´).

TGFBR2+/- mice were obtained from Deltagen, Inc. (San Carlos, CA). A 6.93 kb IRES-lacZ reporter and neomycin resistance cassette (IRES-lacZ-neo) was subcloned into a 4.2 kb fragment isolated from a mouse genomic phage library, such that 106 base pairs coding for the protein were replaced by IRES-LacZ-neo (Supplementary Fig. 4a). The IRES-lacZ-neo cassette was flanked by 1.2 kb of mouse genomic DNA at its 5´ aspect and 3.0 kb of mouse genomic DNA at its 3´ aspect. The targeting vector was linearized and electroporated into 129/OlaHsd mouse embryonic (ES) stem cells. ES cells were selected for G418 resistance, and colonies carrying the homologously integrated neo

DNA were identified by PCR amplification using a 5´ neo-specific primer (5´-GGGATCTTGGCCATGGTAAGCTGAT-3´) paired with a primer located outside the targeting homology arms on the 5´ side (GS1: 5´-GCACAACCTGATCATACTGTATCCA-3´). The homologous recombination event was confirmed on the 3´ side using a 3´ neo-specific primer (5´-ACGTACTCGGATGGAAGCCGGTCTT-3´) paired with a primer located outside the targeting homology arm on the 3´ side. (GS2: 5´-TACCTCATGGCCCATATGACATAAT-3´).

For both *C3AR1-/-* and *TGFBR2+/-* mice, colonies that gave rise to the correct size PCR product were confirmed by Southern blot analysis using a probe adjacent to the 5´ region of homology. The presence of a single neo cassette was confirmed by Southern blot analysis using a neo gene fragment as a probe. Male chimeric mice were generated by injection of the targeted ES cells into C57Bl/6J blastocysts. Chimeric mice were bred with C57Bl/6J mice to produce F1 heterozygotes. Germline transmission was confirmed by PCR analysis. Initial germline heterozygotes were also tested for the homologous recombination event using the primers described above (located outside of the targeting construct) (Supplementary Fig. 3b and 4b). Following confirmation of the targeting event in animals, subsequent genotyping tracked transmission of the targeting construct (Supplementary Fig. 3c-d and 4c-d). F1 heterozygous males and females were mated to produce F2 wild-type, heterozygous and homozygous null mutant animals. Viable embryos for *Tgfbr2-/-* mice were identified, but none survived to birth, so that homozygous knockouts for this gene were determined to be lethal. Mice were backcrossed with C57BL/6J mice and all phenotypic analysis was performed in a hybrid C57Bl/6J/129 background (75%/25%, respectively).

The BAC clone CTD-2339M9, which covers the human *ZFP90* gene sequence with minimal overlaps with neighboring genes, was purchased from Invitrogen

(Carlsbad, CA) and the sequence was confirmed using PCR primers specific to the two BAC ends and the *ZFP90* gene sequence. The clone was cultured in LB medium containing 12.5 μg/ml chloramphenical. BAC DNA was extracted and purified with a Large-Construct kit (QIAGEN, Valencia, CA) and subsequently quantified using pulse field electrophoresis (CHEF-DR^{TM} II Electrophoresis System, BioRad, Hercules, CA) along with High DNA Mass Ladder (Invitrogen, Carlsbad, CA). The purified circular BAC DNA was injected into pronucleus of fertilized FVB eggs and surviving eggs were transferred to pseudopregnant FVB female mice at UCI Transgenic Mouse Facility (http://darwin.bio.uci.edu/%7Etjf/). Three *ZFP90* transgenic founders were identified from DNA isolated from tail biopsies by PCR using BAC-end-specific and gene-specific primers. The integration of the entire BAC sequence was confirmed in all founders. Using RT-PCR, human *ZFP90* transcript was identified from total RNA samples isolated from brain, heart, liver, kidney, and spleen, but not from the adipose tissue (Supplementary Fig. 5).

Control FVB/NJ mice used for comparison to the Zfp90 transgenic were bred in house and at the Jackson Laboratory (Sacramento) from known homozygous parental breeders. All animals were housed 4 per cage at 25°C on a 10-hr dark/14-hr light cycle and had *ad libitum* access to water and to regular rodent chow (Purina diet # 5015).

Supplementary Methods References

1. Basten, C.A., Weir, B.S. & Zeng, Z.B. QTL Cartographer, A Reference Manual and Tutorial for QTL Mapping. (Department of Statistics, North Carolina State University, Raleigh, NC, 1999).

2. Jiang, C. & Zeng, Z.B. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**, 1111-27 (1995).

3. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*, xix, 552 p. (Morgan Kaufmann Publishers, San Mateo, Calif., 1988).

4. Haley, C.S. & Knott, S.A. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315-324 (1992).

5. Shumway, R.H. & Stoffer, D.S. *Time series analysis and its applications*, xiii, 549 p. (Springer, New York, 2000).